Estimating the null distribution for conditional inference and genome-scale screening

David R. Bickel

October 5, 2009

Ottawa Institute of Systems Biology
Department of Biochemistry, Microbiology, and Immunology
Department of Mathematics and Statistics
University of Ottawa
451 Smyth Road
Ottawa, Ontario, K1H 8M5

Abstract

In a novel approach to the multiple testing problem, Efron (2004; 2007) formulated estimators of the distribution of test statistics or nominal p-values under a null distribution suitable for modeling the data of thousands of unaffected genes, non-associated single-nucleotide polymorphisms, or other biological features. Estimators of the null distribution can improve not only the empirical Bayes procedure for which it was originally intended, but also many other multiple comparison procedures. Such estimators serve as the groundwork for the proposed multiple comparison procedure based on a recent frequentist method of minimizing posterior expected loss, exemplified with a non-additive loss function designed for genomic screening rather than for validation.

The merit of estimating the null distribution is examined from the vantage point of conditional inference in the remainder of the paper. In a simulation study of genome-scale multiple testing, conditioning the observed confidence level on the estimated null distribution as an approximate ancillary statistic markedly improved conditional inference. To enable researchers to determine whether to rely on a particular estimated null distribution for inference or decision making, an information-theoretic score is provided that quantifies the benefit of conditioning. As the sum of the degree of ancillarity and the degree of inferential relevance, the score reflects the balance conditioning would strike between the two conflicting terms.

Applications to gene expression microarray data illustrate the methods introduced.

Keywords: ancillarity; attained confidence level; composite hypothesis testing; conditional inference; empirical null distribution; GWA; multiple comparison procedures; observed confidence level; simultaneous inference; simultaneous significance testing; SNP

1 Introduction

1.1 Multiple comparison procedures

1.1.1 Aims of multiple-comparison adjustments

Controversy surrounding whether and how to adjust analysis results for multiple comparisons can be partly resolved by recognizing that a procedure that works well for one purpose is often poorly suited for another since different types of procedures solve distinct statistical problems. Methods of adjustment have been developed to attain three goals, the first two of which optimize some measure of sample space performance:

1. *Adjustment for selection.* The most common concern leading to multiple-comparison adjustments stems from the observation that results can achieve nominal statistical significance because they were selected to do so rather than because of a reproducible effect. Adjustments of this type are usually based on control of a Type I error rate such as a family-wise error rate or a false discovery rate as defined by Benjamini and Hochberg (1995). Dudoit et al. (2003) reviewed several options in the context of gene expression microarray data.

2. *Minimization of a risk function.* Stein (1956) proved that the maximum likelihood estimator (MLE) is inadmissible for estimation of a multivariate normal mean under squared error loss, even in the absence of correlation. Efron and Morris (1973) extended the result by establishing that the MLE is dominated by a wide class of estimators derived via an empirical Bayes approach in which the mean is random. More recently, Ghosh (2006) adjusted p-values for multiple comparisons by minimizing their risk as estimators of a posterior probability. In the presence of genome-scale numbers of comparisons, adjustments based on hierarchical models are often much less extreme than those needed to adjust for selection. For two examples from microarray data analysis, Efron (2008) found that posterior intervals based on a local false discovery rate (LFDR) estimate tend to be substantially narrower than those needed to control the false coverage rate introduced by Benjamini et al. (2005) to account for selection, and an LFDR-based posterior mean has insufficient shrinkage toward the null to adequately correct selection bias (Bickel, 2008a).

3. *Estimation of null or alternative distributions.* Measurements over thousands of biological features available from studies of genome-scale expression and genome-wide association studies have recently enabled estimation of distributions of p-values. Early empirical Bayes methods of estimating the LFDR associated with each null hypothesis employ estimates of the distribution of test statistics or p-values under the alternative hypothesis (e.g., Efron et al., 2001). Efron (2004; 2007a) went further, demonstrating the value of also estimating the distribution of p-values under the null hypothesis provided a sufficiently large number of hypotheses under simultaneous consideration.

While all three aims are relevant to Neyman-Pearson testing, they differ as much in their relevance to Fisherian significance testing as in the procedures they motivate. Mayo and Cox (2006) pointed out that Type I error rate control is appropriate for making series of decisions but not for inductive reasoning, where the inferential evaluation of evidence is of concern apart from loss functions that depend on how that evidence will be used, which, as Fisher (1973, pp. 95-96, 103-106) stressed, might not even be known at the time of data analysis. Likewise, Hill (1990) and Gleser (1990) found optimization over the sample space helpful for making series of decisions rather than for drawing scientific inferences from a particular observed sample. Cox (1958; 2006) noted that selection of a function to optimize is inherently subjective to the extent that different decision makers have different interests. Further, sample space optimality is often achieved at the expense of induction about the parameter given the data at hand; for example, optimal confidence intervals result from systematically stretching them in samples of low variance and reducing them in samples of high variance relative to their conditional counterparts (Cox, 1958; Barnard, 1976; Fraser and Reid, 1990; Fraser, 2004a,b).

The suitability of the methods of both of the first two goals for decision rules as opposed to inductive reasoning is consistent with the observation that control of Type I error rates may be formulated as a minimax problem (e.g., Lehmann 1950; Wald 1961, §1.5), indicating that the second of the above aims generalizes the first. Although corrections in order to account for selection are often applied when it is believed that only a small fraction of null hypotheses are false (Cox, 2006), the methods of controlling a Type I error rate used to make such corrections are framed in terms of rejection decisions and thus may depend on the number of tests conducted, which would not be the case were the degree of correction a function only of prior beliefs. By contrast with the first two aims, the third aim, improved specification of the alternative or null distribution of test statistics, is clearly as important in significance testing as in fixed-level Neyman-Pearson testing. In short, while the first two motivations for multiple comparison procedures address decision-theoretic problems, only the third pertains to significance testing in the sense of impartially weighing evidence without regard to possible consequences of actions that might be taken as a result of the findings.

1.1.2 Estimating the null distribution

Because of its novelty and its potential importance for many frequentist procedures of multiple comparisons, the effect of relying on the following method due to Efron (2004; 2007a; 2007b) of estimating the null distribution will be examined herein. The method rests on the assumption that about 90% or more of a large number of p-values correspond to *unaffected features* and thus have a common distribution called the *true null distribution*. If that distribution is uniform, then the *assumed null distribution* of test statistics with respect to which the p-values were computed is correct.

In order to model the null distribution as a member of the normal family, the p-values are transformed by $\Phi^{-1} : [0, 1] \to \mathbb{R}^1$, the standard normal quantile function. The parameters of that distribution are estimated either by fitting a curve to the central region of a histogram of the transformed p-values (Efron, 2004) or, as used below, by applying a maximum likelihood procedure to a truncated normal distribution (Efron, 2007b). The main justification for both algorithms is that since nearly all p-values are modeled as variates from the true null distribution and since the remaining p-values are considered drawn from a distribution with wider tails, the less extreme p-values better resemble the true null distribution than do those that are more extreme. Since the theoretical null distribution is standard normal in the transformed domain, deviations from the standard normal distribution reflect departures in the less extreme p-values from uniformity in the original domain.

For use in multiple testing, all of the transformed p-values of the data set are treated as test statistics for the derivation of new p-values with respect to the null distribution estimated as described above instead of the assumed null distribution. Such adjusted p-values would be suitable for inductive inference or for decision-theoretic analyses such as those controlling error rates, provided that the true null distribution tends to be closer to the estimated null distribution than it is to the assumed null distribution.

1.2 Overview

The next section presents a confidence-based distribution of a vector parameter in order to unify the present study of null distribution estimation within a single framework. The general framework is then applied to the problem of estimating the null distribution in Section 3.1. Section 3.2 introduces a multiple comparisons procedure for coherent decisions made possible by the confidence-based posterior without recourse to Bayesian or empirical Bayesian models.

Adjusting p-values by the estimated null distribution is interpreted as inference conditional on that estimate in Section 4. The simulation study of Section 4.1 demonstrates that estimation of the null distribution can substantially improve conditional inference even when the assumed null distribution is correct marginal over a precision statistic. Section 4.2 provides a method for determining whether the estimated null distribution is sufficiently ancillary and relevant for effective conditional inference or decision making on the basis of a given data set.

Section 5 concludes with a discussion of the new findings and methods.

2 Statistical framework

2.1 Confidence levels as posterior probabilities

The observed data vector $x \in \Omega$ is modeled as a realization of a random quantity X of distribution P_ξ, a probability distribution on the measurable space (Ω, Σ) that is specified by the *full parameter* $\xi \in \Xi \subseteq \mathbb{R}^d$. Let $\theta = \theta(\xi)$ denote a *parameter of interest* in Θ and $\gamma = \gamma(\xi)$ a *nuisance parameter.*

Definition 1. In addition to the above family of probability measures $\{P_\xi : \xi \in \Xi\}$, consider a family of probability measures $\{P^x : x \in \Omega\}$, each on the space $(\Theta, \mathcal{A})$, and a set $\mathcal{R}(S) = \left\{\hat{\Theta}_{\rho,s(\rho)} : \rho \in [0,1], s \in S\right\}$ of region estimators corresponding to a set S of shape functions, where $\hat{\Theta}_{\rho,s(\rho)} : \Omega \to \mathcal{A}$ for all $\rho \in [0,1]$ and $s \in S$. If, for every $\Theta' \in \mathcal{A}$, $x \in \Omega$, and $\xi \in \Xi$, there exist a coverage rate ρ and shape $s(\rho)$ such that

$$P^x(\Theta') = \rho = P_\xi\left(\theta(\xi) \in \hat{\Theta}_{\rho,s(\rho)}(X)\right) \tag{1}$$

and $\hat{\Theta}_{\rho,s(\rho)}(x) = P^x(\Theta')$, then the probability $P^x(\Theta')$ is the *confidence level* of the hypothesis $\theta(\xi) \in \Theta'$ according to P^x, the *confidence measure* over Θ corresponding to $\mathcal{R}(S)$.

Remark 2. Unless the σ-field $\mathcal{A}$ is Borel, the confidence level of the hypothesis of interest will not necessarily be defined; cf. McCullagh (2004).

Building on work of Efron and Tibshirani (1998) and others, Polansky (2007) used the equivalent of P^x to concisely communicate a distribution of "observed confidence" or "attained confidence" levels for each hypothesis that θ lies in some region Θ'. The decision-theoretic "certainty" interpretation of P^x as a non-Bayesian posterior (Bickel, 2009) serves the same purpose but also ensures the coherence of actions that minimize expected posterior loss. Robinson (1979) also considered interpreting the ratio $\rho/(1-\rho)$ from equation (1) as odds for betting on the hypothesis that $\theta \in \Theta'$.

The posterior distribution need not conform to the Bayes update rule (Bickel, 2009) since decisions that minimize posterior expected loss, or, equivalently, maximize expected utility, are coherent as long as the posterior distribution is some finitely additive probability distribution over parameter space (see, e.g., Anscombe and Aumann, 1963). It follows that an intelligent agent that acts as if $\rho/(1-\rho)$ are fair betting odds for the hypothesis that θ lies in a level-ρ confidence region estimated by some region estimator of exact coverage rate ρ is coherent if and only if its actions minimize expected loss with the expectation value over a confidence measure as the probability distribution defining the expectation value (cf. Bickel, 2009). Minimizing expected loss over the parameter space, whether based on a confidence posterior or on a Bayesian posterior, differs fundamentally from the decision-theoretic approach of Section 1.1 in that the former is optimal given the single sample actually observed whereas the latter is optimal over repeated sampling. Section 3.2 illustrates the minimization of confidence-measure expected loss with an application to screening on the basis of genomics data.

2.2 Confidence levels versus p-values

Whether confidence levels agree with p-values depends on the parameter of interest and on the chosen hypotheses. If θ is a scalar and the null hypothesis is $\theta = \theta'$, the p-values associated with the alternative hypotheses $\theta > \theta'$ and $\theta < \theta'$ are $P^x((-\infty, \theta'))$ and $P^x((\theta', \infty))$, respectively; cf. Schweder and Hjort (2002).

On the other hand, a p-value associated with a two-sided alternative is not typically equal to the confidence level $P^x(\{\theta'\})$. Polansky (2007, pp. 126-128, 216) discusses the tendency of the attained confidence level of a point or simple hypotheses such as $\theta = \theta'$ to vanish in a continuous parameter space. That only a finite number of points in hypothesis space have nonzero confidence is required of any evidence scale that is fractional in the sense that the total strength of evidence over Θ is finite. (Fractional scales enable statements of the form, "the negative, null, and positive hypotheses are 80%, 15%, and 5% supported by the data, respectively.") While the usual two-sided p-value vanishes only for sufficiently large samples, the confidence level $P^x(\{\theta'\})$ typically is 0% even for the smallest samples and thus does not lead to the appearance of a paradox of "over-powered" studies. As a remedy, Hodges and Lehmann (1954) proposed testing an interval hypothesis $\theta \in \Theta'$ defined in terms of scientific significance; in this situation, as with composite hypothesis-testing in general, $P^x(\Theta')$ converges in probability to $1_{\Theta'}(\theta)$ even though the two-sided p-value does not (Bickel, 2009). (Testing a simple null hypothesis against a composite alternative hypothesis yields a similar discrepancy between a two-sided p-value and methods that respect the likelihood principle (Levine, 1996; Bickel, 2008b).)

There are nonetheless situations that, when using p-values for statistical significance, necessitate testing a hypothesis known to be false for all practical purposes. Cox (1977) called a null hypothesis $\theta = \theta'$ *dividing* if it is not considered because it could possibly be approximately true but rather because it lies on the boundary between $\theta < \theta'$ and $\theta > \theta'$, the two hypotheses of genuine interest. For example, a test of equality of means and its associated two-sided p-value often serve the purpose of determining whether there are enough data to determine the direction of the difference when it is known that there is some appreciable difference (Cox, 1977). That goal can be more directly attained by comparing the confidence levels $P^x((-\infty, \theta'))$ and $P^x((\theta', \infty))$. While reporting the ratio or maximum of $P^x((-\infty, \theta'))$ and $P^x((\theta', \infty))$ would summarize the confidence level of each of two regions in a single number, such a number may be more susceptible to misinterpretation than a report of the pair of confidence levels.

2.3 Simultaneous inference

In the typical genome-scale problem, there are d scalar parameters $\theta_1, \theta_2, ..., \theta_d$ and d corresponding observables $X_1, X_2, ..., X_d$, such that $d \geq 1000$ and $\theta_i = \theta_i(\xi)$ is a subparameter of the distribution of X_i, the random quantity of which the observation $x_i \in \Omega_i$ is a realized vector. The ith of the d hypotheses to be simultaneously tested is $\theta_i \in \Theta'_i$ for some Θ'_i in Θ_i, a subset of $\mathbb{R}^1$. Representing numeric tuples under the angular bracket convention to distinguish the open interval (x, y) from the ordered pair $\langle x, y \rangle$, $\theta = \theta(\xi) = \langle \theta_1, \theta_2, \ldots, \theta_d \rangle$ is the d-dimensional subparameter of interest and the joint hypothesis is $\theta(\xi) \in \Theta'$, where $\Theta' = \Theta'_1 \times \Theta'_2 \times \cdots \times \Theta'_d$.

For any $\delta \in \{1, 2, ..., d-1\}$, inference may focus on δ of the scalar parameters as opposed to the entire vector θ. For example, separate consideration of the confidence levels of hypotheses such as $\theta_1 \in \Theta'_1$ or of $\langle \theta_1, \theta_2 \rangle \in \Theta'_1 \times \Theta'_2$ can be informative, especially if d is high. Each of the components of the *focus index* $\iota = \langle i(1), i(2), \ldots, i(\delta) \rangle$ is in $\{1, ..., d\}$ and is unequal to each of its other components. The proper subset $\tilde{\Theta}'_\iota = \Theta'_{i(1)} \times \Theta'_{i(2)} \times \cdots \times \Theta'_{i(\delta)}$ of $\tilde{\Theta}_\iota = \Theta_{i(1)} \times \Theta_{i(2)} \times \cdots \times \Theta_{i(\delta)}$ is defined in order to weigh the evidence for the hypothesis that $\tilde{\theta}_\iota = \left\langle \theta_{i(1)}, \theta_{i(2)}, \ldots, \theta_{i(\delta)} \right\rangle \in \tilde{\Theta}'_\iota$. Setting $\Theta'_\iota = \Theta'_1 \times \Theta'_2 \times \cdots \times \Theta'_d$ such that $\Theta'_j = \Theta_j$ for all $j \notin \{i(1), i(2), \ldots, i(\delta)\}$, define the marginal distribution P^x_ι such that $P^x_\iota\left(\tilde{\Theta}'_\iota\right)$ is equal to the confidence level $P^x(\Theta'_\iota)$. Thus, P^x_ι is a probability measure marginal over all θ_j with $j \notin \{i(1), i(2), \ldots, i(\delta)\}$.

The following lemma streamlines inference focused on whether $\tilde{\theta}_\iota \in \tilde{\Theta}'_\iota$, or, equivalently, $\theta(\xi) \in \Theta'_\iota$, by establishing sufficient conditions for the confidence level marginal over some of the d components of θ to be equal to the parameter coverage probability marginal over the data corresponding to those components.

Lemma 3. *Considering a focus index ι and $\tilde{X}_\iota = \left\langle X_{i(1)}, X_{i(2)}, \ldots, X_{i(\delta)} \right\rangle$, let $\hat{\Theta}^\iota_{\rho,s(\rho)} : \Omega \to \tilde{\mathcal{A}}_\iota$ be the corresponding level-ρ set estimator of some shape parameter $s(\rho)$ defined such that for every $x \in \Omega$, $\hat{\Theta}^\iota_{\rho,s(\rho)}(x)$ is the canonical projection of $\hat{\Theta}_{\rho,s(\rho)}(x)$ from $\mathcal{A}$ to $\tilde{\mathcal{A}}_\iota$, the σ-field of the marginal distribution P^x_ι. If there is a map $\tilde{\Theta}^\iota_{\rho,s(\rho)} : \tilde{\Omega}_\iota \to \tilde{\mathcal{A}}_\iota$ such that $\tilde{\Theta}^\iota_{\rho,s(\rho)}\left(\tilde{X}_\iota\right)$ and $\hat{\Theta}^\iota_{\rho,s(\rho)}(X)$ are identically distributed, then P^x_ι is the confidence measure over $\tilde{\Theta}_\iota$ corresponding to $\left\{\tilde{\Theta}^\iota_{\rho,s(\rho)} : \rho \in [0,1], s \in S\right\}$.*

Proof. By the general definition of confidence level (1),

$$P^x_\iota\left(\tilde{\Theta}'_\iota\right) = P^x(\Theta'_\iota) = P_\xi\left(\theta \in \hat{\Theta}_{\rho,s(\rho)}(X)\right),$$

where the coverage rate ρ and shape parameter $s(\rho)$ are constrained such that $\hat{\Theta}_{\rho,s(\rho)}(x) = \Theta_\iota'$ for the observed value x of random element X. Hence, using A_ι to denote the event that $\theta_j \in \Theta_j'$,

$$P_\iota^x\left(\tilde{\Theta}_\iota'\right) = P_\xi\left(\tilde{\theta}_\iota \in \hat{\Theta}_{\rho,s(\rho)}^\iota(X), A_\iota\right) \tag{2}$$

with the coverage rate ρ and shape parameter $s(\rho)$ restricted such that $\hat{\Theta}_{\rho,s(\rho)}^\iota(x) = \tilde{\Theta}_\iota'$. Considering $j \notin \{i(1), i(2), \dots, i(\delta)\}$, the event A_ι satisfies $P_\xi(A_\iota) = 1$ since $\Theta_j' = \Theta_j$, thereby eliminating A_ι from equation (2). Because $\tilde{\Theta}_{\rho,s(\rho)}^\iota$ exists by assumption, $\tilde{\Theta}_{\rho,s(\rho)}^\iota(\tilde{x}_\iota) = \tilde{\Theta}_\iota'$ results and $\tilde{\Theta}_{\rho,s(\rho)}^\iota\left(\tilde{X}_\iota\right)$ replaces $\hat{\Theta}_{\rho,s(\rho)}^\iota(X)$ in equation (2) since they are identically distributed. Therefore,

$$P_\iota^x\left(\tilde{\Theta}_\iota'\right) = \rho = P_\xi\left(\tilde{\theta}_\iota \in \tilde{\Theta}_{\rho,s(\rho)}^\iota\left(\tilde{X}_\iota\right)\right),$$

where the coverage rate ρ and shape parameter $s(\rho)$ are constrained such that $\tilde{\Theta}_{\rho,s(\rho)}^\iota(\tilde{x}_\iota) = \tilde{\Theta}_\iota'$ for the observed value $\tilde{x}_\iota = \left\langle x_{i(1)}, x_{i(2)}, \dots, x_{i(\delta)}\right\rangle$ of $\tilde{X}_\iota$. □

Conditional independence is sufficient to satisfy the lemma's condition of identically distributed region estimators:

Theorem 4. *If X_i is conditionally independent of X_j and θ_j given θ_i for all $i \neq j$, then, for any focus index ι, there is a map $\tilde{\Theta}_{\rho,s(\rho)}^\iota : \tilde{\Omega}_\iota \to \tilde{\mathcal{A}}_\iota$ such that $\tilde{\Theta}_{\rho,s(\rho)}^\iota(\tilde{x}_\iota) = \hat{\Theta}_{\rho,s(\rho)}^\iota(x)$ with $\tilde{x}_\iota = \left\langle x_{i(1)}, x_{i(2)}, \dots, x_{i(\delta)}\right\rangle$ for every $x \in \Omega$, and the marginal distribution P_ι^x is the confidence measure over $\tilde{\Theta}_\iota$ corresponding to $\left\{\tilde{\Theta}_{\rho,s(\rho)}^\iota : \rho \in [0,1], s \in S\right\}$.*

Proof. By the conditional independence assumption, $\hat{\Theta}_{\rho,s(\rho)}^\iota(X)$ is conditionally independent of θ_j and X_j for all $j \notin \{i(1), i(2), \dots, i(\delta)\}$ given $\tilde{\theta}_\iota$, entailing the existence of a map $\tilde{\Theta}_{\rho,s(\rho)}^\iota : \tilde{\Omega}_\iota \to \tilde{\mathcal{A}}_\iota$ such that $\tilde{\Theta}_{\rho,s(\rho)}^\iota\left(\tilde{X}_\iota\right)$ and $\hat{\Theta}_{\rho,s(\rho)}^\iota(X)$ are identically distributed. Then the above lemma yields the consequent. □

The theorem facilitates inference separately focused on each scalar subparameter θ_i on the basis of the observation that $X_i = x_i \in \Omega_i$:

Corollary 5. *If X_i is conditionally independent of X_j and θ_j given θ_i for all $i \neq j$, then, for any $i \in \{1, 2, \dots, k\}$, the marginal distribution $P_{\langle i\rangle}^x$ is the confidence measure over Θ_i corresponding to some set $\left\{\tilde{\Theta}_{\rho,s(\rho)}^{\langle i\rangle} : \rho \in [0,1], s \in S\right\}$ of interval estimators, each a map $\tilde{\Theta}_{\rho,s(\rho)}^{\langle i\rangle} : \Omega_i \to \tilde{\mathcal{A}}_{\langle i\rangle}$.*

Proof. Under the stated conditions, the theorem entails the existence of a map $\tilde{\Theta}_{\rho,s(\rho)}^{\langle i\rangle} : \tilde{\Omega}_{\langle i\rangle} \to \tilde{\mathcal{A}}_{\langle i\rangle}$ such that $\tilde{\Theta}_{\rho,s(\rho)}^{\langle i\rangle}\left(\tilde{x}_{\langle i\rangle}\right) = \hat{\Theta}_{\rho,s(\rho)}^{\langle i\rangle}(x)$ with $\tilde{x}_{\langle i\rangle} = x_i$ for every $x \in \Omega$ and entails that the marginal distribution $P_{\langle i\rangle}^x$ is the confidence measure over $\tilde{\Theta}_{\langle i\rangle}$ corresponding to $\left\{\tilde{\Theta}_{\rho,s(\rho)}^{\langle i\rangle} : \rho \in [0,1], s \in S\right\}$. □

Remark 6. The applications of Sections 3 and 4 exploit this property in order to draw inferences from the confidence levels $P_{\langle 1\rangle}^x((\inf\Theta_1, \theta')), P_{\langle 2\rangle}^x((\inf\Theta_2, \theta')), \dots, P_{\langle d\rangle}^x((\inf\Theta_d, \theta'))$ of the hypotheses $\theta_1 < \theta'$, $\theta_2 < \theta'$, ..., $\theta_d < \theta'$, respectively, for very large d. Here, $\delta = 1$, each subscript $\langle j\rangle$ is the 1-tuple representation of the vector ι with j as its only component, and θ' is the scalar supremum shared by all d hypotheses.

3 Null estimation for genome-scale screening

3.1 Estimation of the null posterior

In the presence of hundreds or thousands of hypotheses, the novel methodology of Efron (2007a) can improve evidential inference by estimation of the null distribution. While Efron (2007a) originally applied the estimator to effectively condition the LFDR on an estimated distribution of null p-values, he noted that its applications potentially encompass any procedure that depends on the distribution of statistics under the null hypothesis. Indeed, the stochasticity of parameters that enables estimation of the LFDR by the empirical Bayes machinery need not be assumed for the pre-decision purpose of deriving the level of confidence that each gene is differentially expressed. Thus, the methodology of Efron (2007a) outlined in Section in terms of p-values can be appropriated to adjust confidence levels (§2) since $P_{\langle i \rangle}^{x}\left(\left(-\infty,\theta'\right)\right)$, the level of confidence that a scalar subparameter θ_i is less than a given scalar θ', is numerically equal to $p_{\langle i \rangle}^{x}\left(\theta'\right)$, the upper-tailed p-value for the test of the hypothesis that $\theta_i = \theta'$. Specifically, confidence levels are adjusted in this paper according to the estimated confidence measure under the null hypothesis rather than according to an assumed confidence measure under the null hypothesis.

Treating the parameters indicating differential expression as fixed rather than as exchangeable random quantities arguably provides a closer fit to the biological system in the sense that certain genes remain differentially expressed and other genes remain by comparison equivalently expressed across controlled conditions under repeated sampling. While the confidence measure is a probability measure on parameter space, its probabilities are interpreted as a degrees of confidence suitable for coherent decision making (§3.2), not as physical probabilities modeling a frequency of events in the system. The interpretation of parameter randomness in LFDR methods is less clear except when the LFDR is seen as an approximation to a Bayesian posterior probability under a hierarchical model.

Example 7. A tomato development experiment of Alba et al. (2005) yielded $n = 6$ observed ratios of mutant expression to wild-type expression in most of the $d = 13,340$ genes on the microarray with missing data for many genes. For the ith gene, the interest parameter θ_i is the expectation value of X_i, the logarithm of the expression ratio. The hypothesis $\theta_i < 0$ corresponds to downregulation of gene i in the mutant, whereas $\theta_i > 0$ corresponds to upregulation. To obviate estimation of a joint distribution of d parameters, the independence conditions of Corollary 5 are assumed to hold. Also assuming normally distributed X_i, the one-sample t-test gave the upper-tail p-value equal to the confidence level $P_{\langle i \rangle}^{x_i}\left(\mathbb{R}_-\right)$ for each gene. The notation is that of Remark 6, except with the replacement of each x subscript with x_i to emphasize that only the ith observed vector influences the confidence level corresponding to the ith parameter. Efron's (2007b) maximum-likelihood method of estimating the null distribution from a vector of p-values provided the estimated null confidence measure that is very close to the empirical distribution of the data (Fig. 1), which is consistent with but does not imply the truth of all null hypotheses of equivalent expression ($\theta_i = 0$). Using that estimate of the null distribution in place of the uniform distribution corresponding to the Student t distribution of test statistics has the effect of adjusting each confidence level. Since extreme confidence levels are adjusted toward 1/2, the estimated null reduces the confidence level both of genes with large values of $P_{\langle i \rangle}^{x_i}\left(\mathbb{R}_-\right)$ (confidence of the hypothesis $\theta_i < 0$) and of those with large values of $P_{\langle i \rangle}^{x_i}\left(\mathbb{R}_+\right)$ (confidence of the hypothesis $\theta_i > 0$). Fig. 2 displays the effect of this confidence-level adjustment in more detail.

3.2 Genome-scale screening loss

Carlin and Louis (2000, §B.5.2) observed that with a suitable non-additive loss function, optimal decisions in the presence of multiple comparisons can be made on the basis of minimizing posterior expected loss. A simple non-additive loss function is

$$L_{a,c}\left(M,m\right) = cM^{1+a} + m, \tag{3}$$

where M and m are respectively the number of incorrect decisions and the number of non-decisions concerning the d components of θ; $M + m \le d$. The scalars $a \in \mathbb{R}^1$ and $c > 0$ reflect different aspects of risk aversion: a is an acceleration in the sense of quantifying the interactive compounding effect of multiple errors, whereas if $a = 0$, then c is the ratio of the cost of making an incorrect decision to the cost of not making any decision or, equivalently, the benefit of making a correct decision.

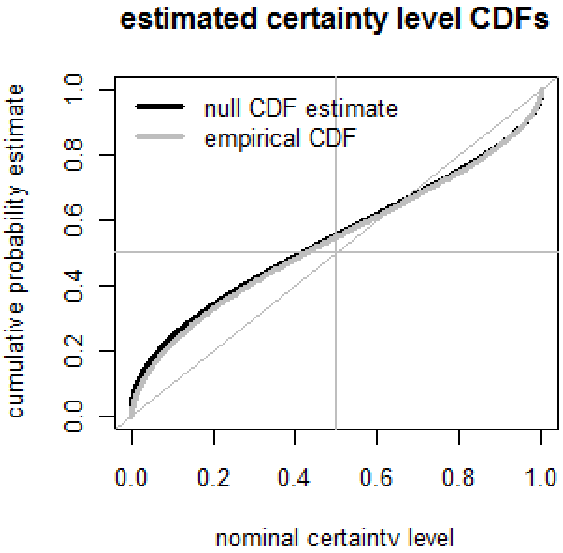


Figure 1: The black curve is the estimated cumulative distribution function (CDF) of the confidence levels under the null distribution, which corresponds to equivalently expressed or unaffected genes; the gray curve is the empirical CDF of all confidence levels, including those of differentially expressed or affected genes. Here, observed confidence coefficients corresponding to hypotheses are interpreted as levels of certainty (§§2.1, 3.2). Departure of the black curve from the diagonal line reflects violation of independence or of the lognormal assumption used to compute the confidence levels. As one-sided p-values, these confidence levels would be uniform under the hypothesis of equivalent expression given the assumptions; i.e., the Φ^{-1}-transformed confidence levels of unaffected genes are assumed to be $\mathrm{N}\left(0, 1^2\right)$, where Φ^{-1} is the standard normal quantile function. The distribution of Φ^{-1}-transformed confidence levels under that null hypothesis was estimated to instead be $\mathrm{N}\left(-0.21, (1.55)^2\right)$. The data set, model, and null distribution estimator are those of Example 7.

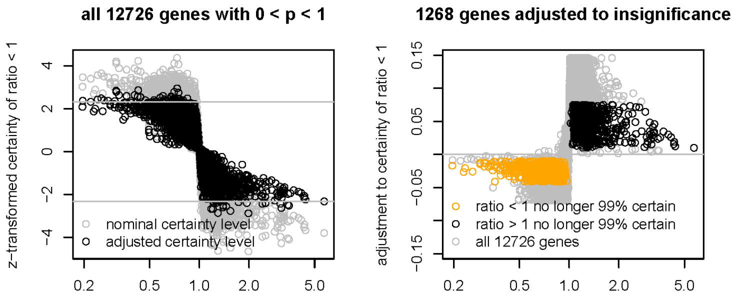


Figure 2: Impact of null estimation on the confidence level as the measure of certainty or statistical significance. The data set, model, and null distribution estimator are those of Example 7 and Fig. 1. *Left panel:* The transformed confidence level $\Phi^{-1}\left(P_{\langle i\rangle}^{x_i}\left(\mathbb{R}_-\right)\right)$ for gene i versus the expression ratio estimated as the geometric sample mean of the observed expression ratio for the same gene. Here, the confidence level $P_{\langle i\rangle}^{x_i}\left(\mathbb{R}_-\right)$ is the degree of certainty of the hypothesis that the mean log-transformed expression ratio is negative or, equivalently, of the hypothesis that the true expression ratio is less than 1. The horizontal lines are drawn at $P_{\langle i\rangle}^{x_i}\left(\mathbb{R}_-\right) = 99\%$ and at $P_{\langle i\rangle}^{x_i}\left(\mathbb{R}_+\right) = 1 - P_{\langle i\rangle}^{x_i}\left(\mathbb{R}_-\right) = 99\%$. Of the original 13,340 genes, 1062 genes have less than the two observations needed for the test statistic and 2 genes have infinite normal-transformed confidence levels and thus are not displayed. Each circle corresponds to a gene, with black for $P_{\langle i\rangle}^{x_i}\left(\mathbb{R}_-;\hat{F}_0\right)$, the confidence level of $\theta_i \in \mathbb{R}_-$ using the estimated null distribution $\hat{F}_0$ and with gray for $P_{\langle i\rangle}^{x_i}\left(\mathbb{R}_-;\tilde{F}_0\right)$, the same except using the assumed null distribution $\tilde{F}_0$. *Right panel:* The difference between $P_{\langle i\rangle}^{x_i}\left(\mathbb{R}_-;\hat{F}_0\right)$ and $P_{\langle i\rangle}^{x_i}\left(\mathbb{R}_-;\tilde{F}_0\right)$ versus the estimated expression ratio. Orange circles represent genes satisfying $P_{\langle i\rangle}^{x_i}\left(\mathbb{R}_-;\tilde{F}_0\right) > 99\%$ but $P_{\langle i\rangle}^{x_i}\left(\mathbb{R}_-;\hat{F}_0\right) \leq 99\%$; black circles represent genes satisfying $P_{\langle i\rangle}^{x_i}\left(\mathbb{R}_+;\tilde{F}_0\right) > 99\%$ but $P_{\langle i\rangle}^{x_i}\left(\mathbb{R}_+;\hat{F}_0\right) \leq 99\%$.

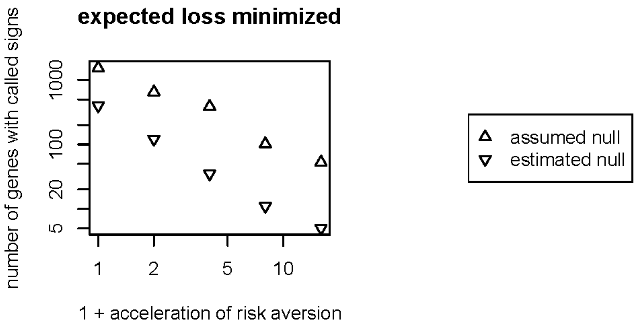


Figure 3: Number $d - m$ of decisions on whether the ith gene is overexpressed ($\theta_i > 0$) or underexpressed ($\theta_i < 0$) plotted against $1 + a$, where a is the degree to which the loss per incorrect decision increases with the number of incorrect decisions (3). The sign call or decision on the direction of regulation for each gene was either made or not made such that the following Monte Carlo approximation to the expected loss $E^x \left(L_{a,9} \left(M, m \right) \right) = \int L_{a,9} \left(M, m \right) dP^x$ was minimized based alternately on the assumed null distribution $\tilde{F}_0$ and on the estimated null distribution $\hat{F}_0$. The kth of the 10^4 values of θ_i was drawn from the frequentist posterior (1) independently for each gene i to compute the correct sign decisions according to the kth realization; such correct decisions yielded M_k and m_k, the number of incorrect sign decisions and the number of non-decisions. The independence of σ-fields corresponding to each gene's scalar component of θ guaranteed by Corollary 5 implies $E^x \left(L_{a,9} \left(M, m \right) \right) \doteq 10^{-4} \sum_{k=1}^{10^4} L_{a,9} \left(M_k, m_k \right)$. The data set, model, and null distribution estimator are those of Example 7 and Figs. 1 and 2.

Bickel (2004) and Müller et al. (2004) applied additive loss ($a = 0$) to decisions of whether or not a biological feature is affected. That special case, however, does not accurately represent the screening purpose of most genome-scale studies, which is to formulate a reasonable number of hypotheses about features for confirmation in a follow-up experiment. More suitable for that goal, $a > 0$ allows generation of hypotheses for at least a few features even on slight evidence without leading to unmanageably high numbers of features even in the presence of decisive evidence.

Fig. 3 displays the result of minimizing such an expected loss with respect to the confidence posterior (1) under the above class of loss functions (3) for decisions on the direction of differential gene expression (Example 7). (Taking the expectation value over the confidence measure rather than over a Bayesian posterior measure was justified in Section 2.1.)

4 Null estimation as conditional inference

4.1 Simulation study

To record the effect of null distribution estimation on inductive inference, a simulation study was conducted with $K = 500$ independent samples each of $d = 10,000$ independent observable vectors, of which 95% correspond to unaffected and 5% to affected features such as genes or single-nucleotide polymorphisms (SNPs). In Example 7, an affected gene is one for which there is differential gene expression between mutant and wild type. Assuming that each scalar parameter θ_i is constrained to lie in the same set Θ_1, the one-sided p-value of each observable is equal to $P_{k,i}^x \left(\left(\inf \Theta_1, \theta' \right) \right)$, the kth confidence level of $\theta_i < \theta'$, the hypothesis that the parameter of interest for the ith observable vector or feature is less than some value θ' dividing two meaningful hypotheses, as discussed in Section 2.2 and illustrated in Fig. 2. (This notation differs from that of Remark 6 in adapting the superscript of the confidence level and from that of Example 7 in dropping the subscript of $x_{k,i}$ for ease of reading.) As $\theta_i = \theta'$ is treated as a null hypothesis for the purpose of estimating or assuming the null distribution, it naturally corresponds an unaffected feature. Each confidence level was generated from Φ, the standard normal CDF, of $Z_{k,i} \sim \mathrm{N} \left(0, \varsigma_k^2 \right)$ for $i \in \{1, \ldots, 9500\}$ or of

$Z \sim \mathrm{N}\left(5\varsigma_k/2, (5\varsigma_k/4)^2\right)$ for $i \in \left\{9501, \dots, 10^4\right\}$. Rather than fixing ς_k at 1 for all k (Efron, 2007a, Fig. 5), ς_k was instead allowed to vary across samples in order to model sample-specific variation that influences the distribution of p-values. For every k in $\{1, \dots, K\}$, $\log \varsigma_k$ is independent and equal to 2/3 with probability 30%, 1 with probability 40%, or 3/2 with probability 30%. Each simulated sample was analyzed with the same maximum-likelihood method of estimating the null distribution used in the above gene expression example, in which the realized value of ς_k was predicted to be about 3/2 (Fig. 1).

Because ς_k is an ancillary statistic in the sense that its distribution is not a function of the parameter and since estimation of the null distribution approximates conditioning the p-values and equivalent confidence levels on the estimated value of ς_k, null estimation is required by the conditionality principle (Cox, 1958), in agreement with the analogy with conditioning on observed row or column totals in contingency tables (Efron, 2007a). See Shi (2008) for further explanation of the relevance of the principle to estimation of the null distribution.

Accordingly, performance of each method of computing confidence levels, whether under the assumed null distribution $\tilde{F}_0$ or estimated null distribution $\hat{F}_0$, was evaluated in terms of the proximity of $P^x_{k,i}\left((\inf \Theta_1, \theta'); F_0\right)$, the confidence level of $\theta_i < \theta'$ for trial k and feature i based on the null hypothesis of distribution $F_0 \in \left\{\hat{F}_0, \tilde{F}_0\right\}$, to $P^x_{k,i}\left((\inf \Theta_1, \theta') | \varsigma_k = \sigma_k\right)$, the corresponding true confidence level conditional on the realized value σ_k of ς_k used to generate the simulated data of trial k. For some $\alpha \in [0, 1]$, the *conservative error* of relying on F_0 as the distribution under the null hypothesis for the kth trial is the average difference in the number of confidence levels incorrectly included in $\mathcal{B} = [\alpha, 1-\alpha]$ and the number incorrectly included in $\bar{\mathcal{B}} = [0,1] \backslash \mathcal{B}$:

$$\sum_{i \in \mathcal{I}} \frac{1_{\mathcal{B}}\left(P^x_{k,i}\left(\Theta'_1; F_0\right)\right) 1_{\bar{\mathcal{B}}}\left(P^x_{k,i}\left(\Theta'_1 | \sigma_k\right)\right) - 1_{\mathcal{B}}\left(P^x_{k,i}\left(\Theta'_1 | \sigma_k\right)\right) 1_{\bar{\mathcal{B}}}\left(P^x_{k,i}\left(\Theta'_1; F_0\right)\right)}{|\mathcal{I}|}, \tag{4}$$

where $\Theta'_1 = (\inf \Theta_1, \theta')$ and where $\mathcal{I} = \{1, \dots, 9500\}$ for the unaffected features or $\mathcal{I} = \left\{9501, \dots, 10^4\right\}$ for the affected features. Here, $\alpha = 1\%$ to quantify performance near confidence values relevant to the inference problem of interpreting the value of $P^x_{k,i}\left((\inf \Theta_1, \theta'); F_0\right)$ as a degree of evidential support for $\theta_i < \theta'$. Values of the conservatism (4) for the simulation study described above appear in Fig. 4.

To determine the effect of analyzing confidence levels that are valid marginal (unconditional) p-values for the mixture distribution, the confidence levels valid given $\varsigma_k = 1$ were transformed such that those corresponding to unaffected features are tail-area probabilities under the marginal null distribution:

$$P_{\theta'}\left(Z_{k,i} < z_{k,i}\right) \quad = \sum_{\sigma \in \{2/3, 1, 3/2\}} P\left(\varsigma_k = \sigma\right) P_{\theta'}\left(Z_{k,i} < z_{k,i} | \varsigma_k = \sigma\right),$$

where $\Phi\left(z_{k,i}\right)$ or $P_{\theta'}\left(Z_{k,i} < z_{k,i}\right)$ is the observed confidence level of $\theta_{k,i} < \theta'$ before or after transformation, respectively. Fig. 5 displays the results.

4.2 Merit of estimating the null distribution

While the degree of undesirable conservatism illustrates the potential benefit of null estimation (§4.1), it does not provide case-specific guidance on whether to estimate the null distribution for a given data set generated by an unknown distribution. Framing the estimated null distribution as a conditioning statistic makes such guidance available from an adaptation of a general measure (Lloyd, 1992) that quantifies the benefit of conditioning inference on a given statistic. Since an approximately ancillary statistic can be much more relevant for inference than an exactly ancillary statistic, Lloyd (1992) quantified the benefit of conditioning on a statistic by the sum of its degree of ancillarity and its degree of relevance, each degree defined in terms of observed Fisher information. To assess the benefit of conditioning inference on the estimated null distribution, the ancillarity and relevance are instead measured in terms of some nonnegative divergence or *relative information*

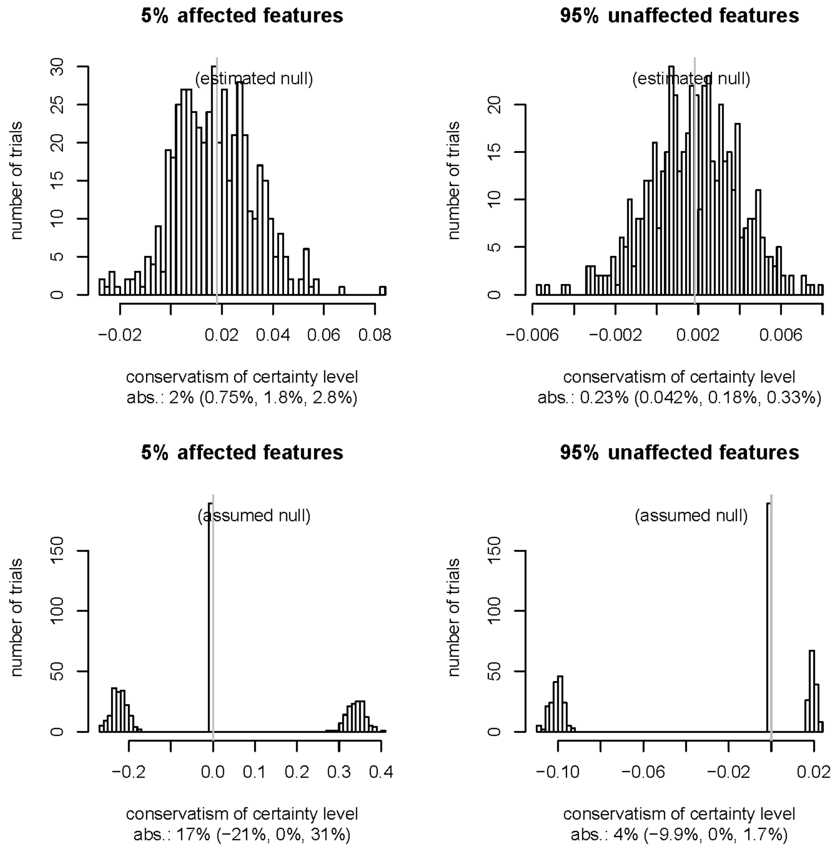
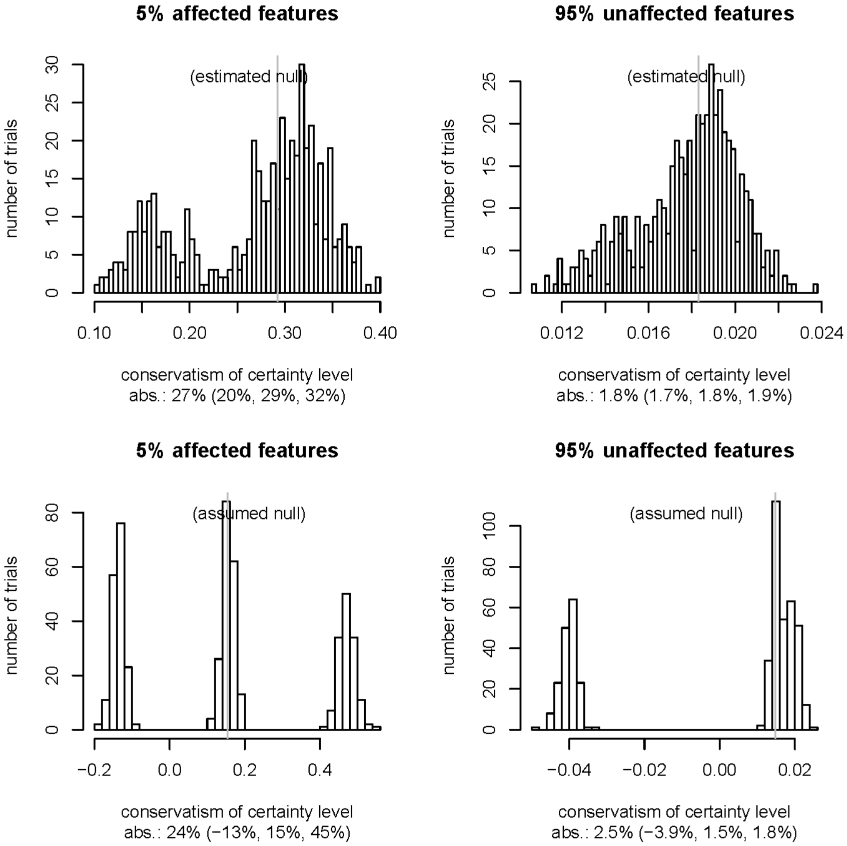


Figure 4: Conservative error (4) when the assumed null distribution is equal to the true null distribution conditional on the most common value of the precision statistic $(\varsigma_k = 1)$. The null distribution F_0 is the estimated distribution $\hat{F}_0$ in the top two plots and the assumed distribution $\tilde{F}_0$ in the bottom two plots. The two plots on the left and right give the errors averaged over the 500 false and the 9500 true null hypotheses, respectively.



5% affected features

95% unaffected features

5% affected features

95% unaffected features

Figure 5: Conservative error (4) when the assumed null distribution is equal to the true null distribution marginal over the distribution of precision statistic ς_k. The four plots have the same arrangement as those of Fig. 4.

$I(F||G)$ between distributions F and G as follows. The ancillarity of the estimated distribution $\hat{F}_0$ for d_1 affected features is the extent to which the parameter of interest is independent of the estimate:

$$A(d_1) = -I\left(\hat{F}_0^{d_1}||\hat{F}_0\right). \tag{5}$$

Here, $\hat{F}_0^{d_1}$ represents the estimated null distribution with its d_1 affected features replaced with unaffected features. More precisely, $\hat{F}_0^{d_1}$ is the estimate of the null distribution obtained by replacing each of the d_1 confidence levels farthest from 0.5 with $(r - 1/2)/d$, the expected order statistic under the assumed null distribution, where r is the rank of the distance of the replaced confidence level from 0.5. Exact ancillarity, $A(d_1) = 0$, thus results only when $\hat{F}_0^{d_1} = \hat{F}_0$, which holds approximately for all d_1 if $\hat{F}_0$ is close to the assumed null distribution. Conditioning on a null distribution estimate is effective to the extent that its relevance,

$$R = I\left(\hat{F}_0||\tilde{F}_0\right), \tag{6}$$

is higher than its *nonancillarity*, $I\left(\hat{F}_0^{d_1}||\hat{F}_0\right)$.

The importance of tail probabilities in statistical inference calls for a measure of divergence $I(F||G)$ between distributions F and G with more tail dependence than the Kullback-Leibler divergence. The Rényi divergence $I_q(F||G)$ of order $q \in (0,1)$ satisfies this requirement, and $I_{1/2}(F||G)$ has proved effective in signal processing as a compromise between the divergence with the most extreme dependence on improbable events ($\lim_{q\to 0} I_q(F||G)$) and the Kullback-Leibler divergence ($\lim_{q\to 1} I_q(F||G)$). Another advantage of $q = 1/2$ is that the commutivity property $I_q(F||G) = I_q(G||F)$ holds only for that order. The notation presents $I_q(F||G)$ as the order-q *information* gained by replacing G with F (Rényi, 1970, §9.8). Since the random variables of the assumed and estimated null distributions are p-values or confidence levels transformed by Φ^{-1} (Fig. 1) and since both distributions are normal, the relative information of order 1/2 is simply

$$I_{1/2}(F||G) = -2\log_2\left(\frac{(\mu_F - \mu_G)^2}{4(\sigma_F^2 + \sigma_G^2)} + \frac{1}{2}\ln\left(\frac{\sigma_F^2 + \sigma_G^2}{2\sigma_F\sigma_G}\right)\right)$$

with $F = \mathrm{N}(\mu_F, \sigma_F^2)$ and $G = \mathrm{N}(\mu_G, \sigma_G^2)$.

Assembling the above elements, the net *inferential benefit* of estimating the null distribution is

$$B(d_1) = A(d_1) + R = I_{1/2}\left(\hat{F}_0||\tilde{F}_0\right) - I_{1/2}\left(\hat{F}_0^{d_1}||\hat{F}_0\right) \tag{7}$$

if there are d_1 affected features, where $\tilde{F}_0 = \mathrm{N}(0,1)$ and where the ancillarity $A(d_1)$ and relevance R are given by equations (5) and (6) with $I = I_{1/2}$. Basing inference on the estimated null distribution is effective to the extent that $B(d_1) > 0$. Fig. 6 uses the gene expression data to illustrate the use of $B(d_1)$ to determine whether to rely on the estimated null distribution $\hat{F}_0$ or on the assumed null distribution $\tilde{F}_0$ for inference.

5 Discussion

Whereas most adjustments for multiple comparisons are aimed at minimizing net loss incurred over a series of decisions optimized over the sample space rather than at weighing evidence in a particular data set for a hypothesis, adjustments resulting from estimation of the distribution of test statistics under the null hypothesis are appropriate for all forms of frequentist hypothesis testing (§1.1). A form seldom considered in non-Bayesian contexts is that of making coherent decisions by minimizing loss averaged over the parameter space. Taking a step toward filling this gap, Section 3.2 provides a loss function suitable for genome-scale screening rather than for confirmatory testing and illustrates its application to the detecting evidence of gene upregulation or downregulation in microarray data.

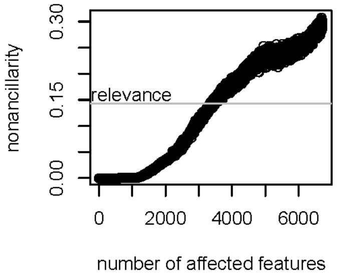


Figure 6: The nonancillarity $-A(d_1)$ versus the hypothetical number k_1 of affected features. The gray horizontal line is the relevance R of null estimation and thus indicates the point at which conditioning on the estimate goes from beneficial ($|A(d_1)| < R$) to deleterious ($|A(d_1)| > R$) according to equation (7). The data set, model, and null distribution estimator are those of Example 7 and Figs. 1, 2, and 3.

Simulations measured the extent to which estimating the null distribution improves conditional inference in an extreme multiple-comparisons setting such as that of finding evidence for differential gene expression in microarray measurements (§4.1). While confidence levels of evidence tended to err on the conservative side under both the estimated and assumed null distributions, conservative error quantified by numbers of confidence levels in $[1\%, 99\%]$ compared to the confidence levels conditional on the precision statistic ς_k was excessive under the assumed null but negligible under the estimated null (Fig. 4). (Since the same pattern of relative conditional performance was obtained by more realistically setting $\log \varsigma_k$ equal to a variate that is independent and uniformly distributed between $\log(1/2)$ and $\log(2)$, those results were not displayed.) Due to the heavy tails of the marginal distribution of pre-transformed confidence levels under the null hypothesis, transforming them to satisfy that distribution under the assumed null increased their conditional conservatism, resulting in about the same performance of estimated and assumed null distributions with respect to the affected features. The case of the unaffected features is more interesting: the assumed null distribution, which after the transformation is marginally exact and hence valid for Neyman-Pearson hypothesis testing, incurs 35% more conservative error than the estimated null distribution (Fig. 5). Thus, the use of the marginal null distribution in place of $\mathrm{N}(0,1)$, the distribution conditional on the central component of the mixture, substantially increases conservative error irrespective of whether the null is assumed or estimated. These results suggest that confidence levels better serve inductive inference when derived from a plausible conditional null distribution than from the marginal distribution even though the latter conforms to the Neyman-Pearson standard. This recommendation reinforces the conditionality principle, which is appropriate for the inferential goal of significance testing as opposed to the various decision-theoretic motivations behind Neyman-Pearson testing (§1.1).

Since the findings of the simulation study do not guarantee the effectiveness of an estimated null distribution $\hat{F}_0$ over the assumed null distribution $\tilde{F}_0$, Section 4.2 gave an information-theoretic score for determining whether to depend on $\hat{F}_0$ in place of$\tilde{F}_0$ for inference on the basis of a particular data set. The score serves as a tool for discovering whether the ancillarity and inferential relevance of $\hat{F}_0$ call for its use in inference and decision making.

6 Acknowledgments

This research was partially supported by the Faculty of Medicine of the University of Ottawa and by Agriculture and Agri-Food Canada. I thank Xuemei Tang for providing the fruit development microarray data. The *Biobase* (Gentleman et al., 2004) and *locfdr* (Efron, 2007b) packages of R (R Development Core Team, 2008) facilitated the computational work.

References

Alba, R., Payton, P., Fei, Z., McQuinn, R., Debbie, P., Martin, G. B., Tanksley, S. D., Giovannoni, J. J., 2005. Transcriptome and selected metabolite analyses reveal multiple points of ethylene control during tomato fruit development. Plant Cell 17 (11), 2954–2965.

Anscombe, F. J., Aumann, R. J., Mar. 1963. A definition of subjective probability. The Annals of Mathematical Statistics 34 (1), 199–205.

Barnard, G. A., 1976. Conditional inference is not inefficient. Scandinavian Journal of Statistics 3 (3), 132–134.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B 57, 289–300.

Benjamini, Y., Yekutieli, D., Edwards, D., Shaffer, J. P., Tamhane, A. C., Westfall, P. H., Holland, B., Benjamini, Y., Yekutieli, D., 2005. False discovery rate-adjusted multiple confidence intervals for selected parameters. Journal of the American Statistical Association 100 (469), 71–93.

Bickel, D. R., 2004. Error-rate and decision-theoretic methods of multiple testing: Which genes have high objective probabilities of differential expression? Statistical Applications in Genetics and Molecular Biology 3 (1), 8.

Bickel, D. R., 2008a. Correcting the estimated level of differential expression for gene selection bias: Application to a microarray study. Statistical Applications in Genetics and Molecular Biology 7 (1), 10.

Bickel, D. R., 2008b. The strength of statistical evidence for composite hypotheses with an application to multiple comparisons. Technical Report, Ottawa Institute of Systems Biology, COBRA Preprint Series, Article 49, available at tinyurl.com/7yaysp.

Bickel, D. R., 2009. Coherent frequentism. Technical Report, Ottawa Institute of Systems Biology, arXiv.org e-print 0907.0139.

Carlin, B. P., Louis, T. A., June 2000. Bayes and Empirical Bayes Methods for Data Analysis, Second Edition, 2nd Edition. Chapman & Hall/CRC, New York.

Cox, D. R., 1958. Some problems connected with statistical inference. The Annals of Mathematical Statistics 29 (2), 357–372.

Cox, D. R., 1977. The role of significance tests. Scandinavian Journal of Statistics 4, 49–70.

Cox, D. R., 2006. Principles of Statistical Inference. Cambridge University Press, Cambridge.

Dudoit, S., Shaffer, J. P., Boldrick, J. C., 2003. Multiple hypothesis testing in microarray experiments. Statistical Science 18 (1), 71–103.

Efron, B., 2004. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. Journal of the American Statistical Association 99 (465), 96–104.

Efron, B., 2007a. Correlation and large-scale simultaneous significance testing. Journal of the American Statistical Association 102 (477), 93–103.

Efron, B., 2007b. Size, power and false discovery rates. Annals of Statistics 35, 1351–1377.

Efron, B., 2008. Microarrays, empirical bayes and the two-groups model. Statistical Science 23 (1), 1–22.

Efron, B., Morris, C., 1973. Stein's estimation rule and its competitors–an empirical bayes approach. Journal of the American Statistical Association 68 (341), 117–130.

Efron, B., Tibshirani, R., 1998. The problem of regions. Annals of Statistics 26 (5), 1687–1718.

Efron, B., Tibshirani, R., Storey, J. D., Tusher, V., 2001. Empirical bayes analysis of a microarray experiment. J. Am. Stat. Assoc. 96 (456), 1151–1160.

Fisher, R. A., 1973. Statistical Methods and Scientific Inference. Hafner Press, New York.

Fraser, D. A. S., 2004a. Ancillaries and conditional inference. Statistical Science 19 (2), 333–351.

Fraser, D. A. S., 2004b. [ancillaries and conditional inference]: Rejoinder. Statistical Science 19 (2), 363–369.

Fraser, D. A. S., Reid, N., Jun. 1990. Discussion: An ancillarity paradox which appears in multiple linear regression. The Annals of Statistics 18 (2), 503–507.

Gentleman, R. C., Carey, V. J., and, D. M. B., 2004. Bioconductor: Open software development for computational biology and bioinformatics. Genome biology 5, R80.

Ghosh, D., 2006. Shrunken p-values for assessing differential expression with applications to genomic data analysis. Biometrics 62 (4), 1099–1106.

Gleser, L. J., Jun. 1990. Discussion: An ancillarity paradox which appears in multiple linear regression. The Annals of Statistics 18 (2), 507–513.

Hill, B. M., Jun. 1990. Discussion: An ancillarity paradox which appears in multiple linear regression. The Annals of Statistics 18 (2), 513–523.

Hodges, J. L., J., Lehmann, E. L., 1954. Testing the approximate validity of statistical hypotheses. Journal of the Royal Statistical Society. Series B (Methodological) 16 (2), 261–268.

Lehmann, E. L., Mar. 1950. Some principles of the theory of testing hypotheses. The Annals of Mathematical Statistics 21 (1), 1–26.

Levine, R.A., C. G., 1996. Convergence of posterior odds. Journal of Statistical Planning and Inference 55 (3), 331–344.

Lloyd, C., 1992. Effective conditioning. Austral. J. Statist. 34, 241–260.

Mayo, D. G., Cox, D. R., 2006. Frequentist statistics as a theory of inductive inference. IMS Lecture Notes - Monograph Series, The Second Erich L. Lehmann Symposium - Optimality.

McCullagh, P., 2004. Fiducial prediction. Technical Report, University of Chicago.

Müller, P., Parmigiani, G., Robert, C., Rousseau, J., 2004. Optimal sample size for multiple testing: the case of gene expression microarrays. Journal of the American Statistical Association 99, 990–1001.

Polansky, A. M., 2007. Observed Confidence Levels: Theory and Application. Chapman and Hall, New York.

R Development Core Team, 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rényi, A., 1970. Probability Theory. North-Holland, Amsterdam.

Robinson, G. K., 1979. Conditional properties of statistical procedures. The Annals of Statistics 7 (4), 742–755.

Schweder, T., Hjort, N. L., 2002. Confidence and likelihood. Scandinavian Journal of Statistics 29 (2), 309–332.

Shi, J., L. D. W. A., 2008. Significance levels for studies with correlated test statistics. Biostatistics 9 (3), 458–466.

Stein, C., 1956. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability 1, 197–206.

Wald, A., 1961. Statistical Decision Functions. John Wiley and Sons, New York.